

Artificial intelligence model for the prediction of cannabis addiction

Abdelilah Elhachimi¹, Mohamed Eddabbah², Abdelhafid Benksim³, Hamid Ibannid⁴,
Mohamed Cherkaoui¹

¹Department of Biology, Cadi Ayyad University, Marrakech, Morocco

²Department of Mathematics and Computer Science, The Higher School of Technology of Essaouira, Cadi Ayyad University, Marrakech, Morocco

³Department of Nursing, Institute of Nursing Professions and Healthcare Techniques (ISPITS), Marrakech, Morocco

⁴National Association of Drug-Risk Reduction (DRR), Marrakech, Morocco

Article Info

Article history:

Received Aug 22, 2024

Revised Nov 12, 2024

Accepted Jan 14, 2025

Keywords:

Cannabis addiction
Combined algorithm
K-means clustering
Linear regression
Machine learning

ABSTRACT

A novel approach for predicting cannabis addiction has been introduced by integrating combined machine learning (ML) algorithms, specifically K-means clustering and linear regression (LR). The study, conducted in Marrakech, Morocco, at a center linked to the National Association for drug-risk reduction (DRR), involved 146 participants. Among those with prior cannabis use, one subgroup included passive users, while another exhibited cannabis dependence. The research utilized features derived from patient data, emphasizing psycho-cognitive state, addiction status, and socio-demographic factors. The goal was to evaluate the effectiveness of the combined ML algorithms (K-means + LR) in distinguishing between addicted and non-addicted individuals using real-world data from a primary care addiction center. The findings indicate that the proposed method delivers balanced results, achieving an overall accuracy of 70%, a sensitivity of 65%, and a specificity of 86%. These results are particularly noteworthy when compared to other ML studies in addiction research. The combined algorithm demonstrates promising potential with competitive accuracy and high specificity. Further efforts to improve sensitivity and validate the model in diverse settings will be essential for advancing predictive modeling in this field. Our findings contribute to existing research by developing simple and effective tools for early detection of cannabis addiction, potentially aiding in the creation of preventive and therapeutic strategies to reduce its prevalence.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Mohamed Eddabbah

Department of Mathematics and Computer Science, The Higher School of Technology of Essaouira

Cadi Ayyad University

Marrakech, Morocco

Email: eddabbah@gmail.com or m.eddabbah@uca.ac.ma

1. INTRODUCTION

The United Nations Office on Drugs and Crime (UNODC) report revealed that cannabis was the most widely consumed substance in 2020, with 209 million users, a 23% increase from 2010 to 2020. Among the 38.6 million individuals suffering from drug use disorders, 40% were affected by cannabis-related disorders. Furthermore, this report noted that 4% of the 494,000 drug-related deaths were linked to cannabis, translating to approximately 19,760 cannabis-related deaths [1]. Cannabis use has numerous negative impacts on individuals, communities, and economies. Specifically, users face health complications, legal issues, social and psychological stigma, and reduced income [2], [3]. Cannabis is the most commonly used drug

among young people [4]. In fact, adolescents are particularly susceptible to mental health issues related to cannabis use, which can lead to cognitive impairments, distraction, and attention difficulties [5]. Indeed, adolescence is a critical period for the development of cognitive abilities, emotional regulation, and sensory-motor functions [6], [7]. Additionally, cognitive impairments resulting from early cannabis use among adolescents are often associated with decreased academic performance and a higher likelihood of dropping out of school [7]. Therefore, early screening for cannabis addiction is crucial to mitigating its adverse effects, considering the significant costs it imposes on individuals, economies, and societies.

Screening methods for cannabis use are mainly divided into biological tests and clinical assessments. On the one hand, biological methods detect cannabis in bodily fluids like urine, blood, saliva, and hair [8]. However, these tests have limitations, including specificity to certain reference molecules, which may allow some cannabis-related compounds to go undetected [8]. Moreover, a positive biological test only confirms cannabis use, not addiction [9]. On the other hand, clinical evaluations emphasize addiction by examining consumption patterns and psychological conditions [10]. However, these questionnaires may introduce biases such as underreporting due to shame or denial [11], [12]. Moreover, the majority of these questionnaires lack alignment with the criteria outlined in the fifth edition of the diagnostic and statistical manual of mental disorders (DSM-5). Consequently, clinical diagnostics based on these tools should be performed under the supervision of an addiction specialist to ensure the accuracy and reliability of the screening process [10].

Recently, the integration of machine learning (ML) and artificial intelligence (AI) techniques in healthcare had a positive impact on almost every aspect of healthcare, from disease detection to the prediction of its evolution and prevention [13]. However, although it offers benefits, the utilization of ML in addiction research remains limited [14]. Besides, one of the most common applications of ML is making predictions. For instance, it can be used to differentiate individual characteristics (e.g., those with and without binge drinking behavior) or to predict events (e.g., an opioid overdose) [15], [16].

It's typical to categorize ML approaches into two primary groups: supervised learning and unsupervised learning. The use of supervised learning for addiction problems allows to identify the population at risk, extract the most relevant variables associated with substance use disorders (SUD), and to differentiate the population with or without SUD. While unsupervised learning methods of discovering emerging relationships and groupings within the data without any predefined target [17]. K-means clustering stands out as a prevalent algorithm in unsupervised ML, frequently utilized, for example, to categorize individuals with SUD based on comparable psychosocial or clinical characteristics [18]. In classification problems, the response is typically organized into categories (e.g., the occurrence of an opioid overdose) [16]. Another facet of ML is deep learning, whose applications in addictionology enable highly accurate classification in detecting key indicators for individuals addicted to opioids, particularly those with long-term use [19].

ML holds the capability to address a diverse array of challenges in addictology. Indeed, using a ML framework to forecast the efficacy of treatment for SUD [17]. Additionally, the use of ML and functional magnetic resonance imaging data as a potential biomarker for the identification of cocaine dependence [20]. Besides, the use of ML can help accurately identify high-risk youth and young adults in need of SUD prevention [18]. Lastly, ML can aid in the discovery of the genetic underpinnings of addiction [21]. Recently, a study presented a method for screening cannabis dependence by combining ML with psychological and cognitive assessment tests, highlighting the success of the support vector machine (SVM) model in identifying cannabis addiction [22].

Undoubtedly, these studies are very promising in regards to improving the field of psycho-addictology. However, these studies are limited by four aspects, namely: i) the large number of used variables, ii) the proportionality of the classes, iii) the choice of the metrics for the assessments of the models, and vi) the use of baseline ML algorithms without data preprocessing and without algorithm adaptation to the context. This study aimed to test the ability of a combined ML algorithm (K-means + linear regression (LR)) to differentiate between addicted and non-addicted patients using real data collected from a primary care center for addiction. The gathered data focuses primarily on the psycho-cognitive state, drug addiction status, and sociodemographic background of each patient. The objective of this study is to develop a reliable, cost-efficient, and accurate ML-based screening and classification tool for prediction of cannabis addiction using approved, easy-to-use objective clinical tests. This tool is designed to be accessible even to non-practitioners, addressing understaffing challenges and aiding in the implementation of preventive measures.

2. METHOD

2.1. Study design

This study aimed to predict cannabis dependence among adolescents using ML techniques. The prediction is based on data from a cross-sectional study conducted in a primary healthcare center specializing in addictology. The data from study was mainly aimed at: i) describe the participants sociodemographic

profiles, ii) analyze the characteristics of cannabis dependence, iii) assess the presence of cognitive disorders, anxiety, and depression, and iv) evaluate sleep quality.

2.2. Population study

A population of 146 participants from both genders, aged between 13 to 25 years old ($M=20.4$; $SD=2.7$), and who agreed to participate in the study. This sample comprises two groups: 73 participants clinically identified as cannabis addicts, and 73 control patients without cannabis addiction. The exclusion criteria included; i) patients who refused to participate in the study, ii) patients who were unable to answer the questions and who had severe behavioral problems, and iii) patients who were addicted to several psychotropic substances apart from tobacco. Samples included in this study were collected through simple random sampling from all patients meeting the inclusion and exclusion criteria mentioned above.

2.3. Data collection and statistical analysis

Following a consultation with the Marrakech DRR center's psychiatrist-addictologist, an anonymous questionnaire is used to conduct an interview aiming at examining various sociodemographic, cannabis addiction, psychological, cognitive, and sleep quality characteristics. Considering the neuropsychological effects of cannabis consumption, specific tools are utilized to assess the psychological and neurocognitive impacts associated with cannabis dependence. These instruments aim to evaluate cannabis use disorder (CUD), problematic cannabis use, anxiety and depression levels, sleep quality, and cognitive function.

This study was conducted at the Marrakech DRR-Maroc center. The study was authorized by the regional health directorate. The procedures were conducted in accordance with the guidelines of the declaration of Helsinki. All participants were informed prior to data collection about the purpose of the study. Also, we have obtained the written informed consent of each participant and/or their legal representative. The statistical analysis conducted in this research as well as ML modeling were both performed using the Python language and the scikit-learn package.

2.4. Study variables

2.4.1. Socio-demographic information

The demographic data included information on gender, age, occupation, marital status, geographical origin, education level, and professional activity, offering a comprehensive understanding of the population's socio-economic background. It provides key insights into the diverse characteristics and challenges faced by different groups. This information helps identify factors that influence behaviors and decisions within the population.

2.4.2. Characteristic of cannabis addiction

The study assessed the addictive traits of cannabis use, including verifying the patient's addiction, age at first use, duration of use, and family history of cannabis use. Additionally, the cannabis abuse screening test (CAST) and DSM-5 CUD tools were employed to measure problematic cannabis use and the level of addiction, respectively. The DSM-5 CUD is a set of guidelines issued by the American psychiatric association to characterize problematic cannabis use in cognitive-behavioral, psychological, and environmental terms. A score under 2 indicates no addiction, a score between 2 and 3 indicates mild addiction, a score between 4 and 5 indicates moderate addiction, and a score above 6 indicates severe addiction [23]. The CAST is a 6-item tool specifically developed to identify patterns of cannabis abuse in adolescents and young adults, focusing on challenges in controlling use and the negative impacts on health or social relationships [24]. A score below 3 suggests no risk of addiction. A score between 3 and less than 7 indicates a low risk of addiction, while a score of 7 or higher signifies a high risk of addiction [25].

2.4.3. Anxiety/depression levels

The presence of anxiety and depression in adolescents was assessed using the hospital anxiety and depression scale (HAD). It is a clinical screening tool for identifying anxiety and depressive disorders. The scale consists of 14 questions, 7 of which are related to anxiety, and 7 to depression. Each question is scored between 0 and 3, and the total score for each dimension ranges from 0 to 21. A score greater than 8 in either dimension indicates a significant presence of anxiety or depression [26].

2.4.4. The psychiatric profiles of patients and their families

Working alongside the healthcare team, we gathered detailed clinical information for each patient, including both personal and family histories of psychiatric disorders. This data offers a thorough perspective on mental health concerns and supports a deeper understanding of the patients' psychological backgrounds. By incorporating these insights, we strive to build a more complete profile for each patient, improving the precision of diagnostic evaluations.

2.4.5. Cognitive status

Cognitive function was evaluated using the Montreal Cognitive Assessment (MoCA) test. It is a clinical neuropsychological test used to assess cognitive impairment. It consists of assessments of executive, visio-spatial, denominative, memory, attention, language, abstraction, delayed recall, and orientation functions. The highest possible score is 30 points. When the score does not exceed the threshold of 26, the patient is identified as having a cognitive impairment [22].

2.4.6. Sleep quality

Sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI). The PSQI is a standardized instrument designed to evaluate sleep efficiency and overall sleep quality. The PSQI includes 11 items, each rated on a scale from 0 to 3, with a total score ranging from 0 to 21. The global PSQI score is calculated by summing the scores of seven components, with a score above 5 indicating poor sleep quality [27].

2.5. Machine learning models

ML refers to a collection of techniques that enable machines to learn, model, and interpret complex datasets without direct human intervention [17]. Figure 1 shows the proposed methodology for this work. After the questionnaire's data was collected, the data underwent conversion during the preprocessing step to ensure it was of high quality and suitable for analysis. Indeed, the collected data will be processed using min-max normalization. This method is effective for managing non-uniform data that falls outside the 0-1 range. It is favored because it ensures that the data remains balanced before and after normalization [28]. The min-max normalization process is illustrated, and the formula is as:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

X' represents the min-max value, X is the value to be normalized, X_{\min} is the lowest value of the overall data, and X_{\max} is the highest value of the entire data.

Next, the feature extraction process will be applied to the data in order to obtain a new, informative, and compact set of features. After that, the gathered data will be processed using the clustering technique. The data of the study population will be clustered into addict and non-addict groups using the K-means algorithm. Each class will be modeled using LR, which will provide a more significant and consistent data display. Then, to improve the reliability of the prediction result and to enhance the prediction model's accuracy, the new patient data will be added to the database following diagnosis.

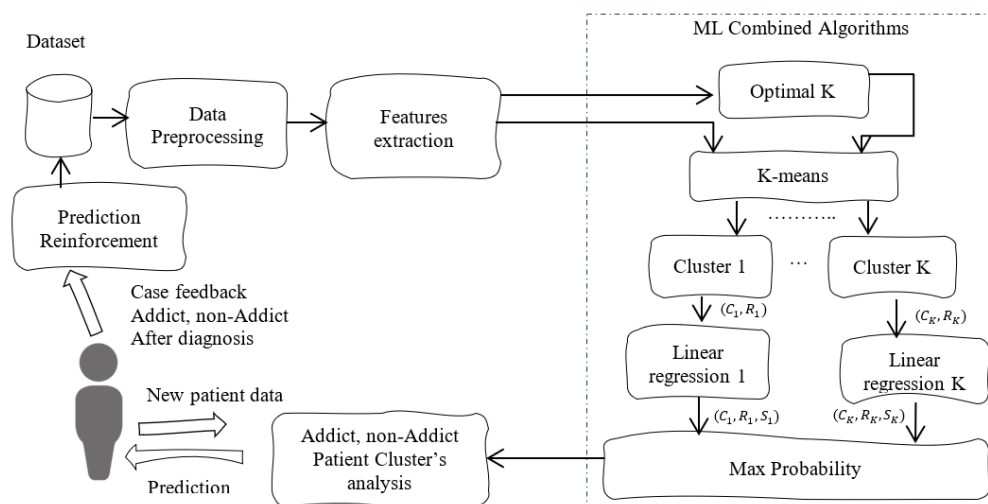


Figure 1. Proposed model for cannabis addiction prediction

Figure 2 represents a ML pipeline for classifying individuals into "Addict" or "Non-Addict" categories. Indeed, the dataset is split into two portions: the 80% portion is likely used for training, while the 20% is likely used for testing or validation. After that, the features are fed into a ML combination algorithm. The ML model outputs classifications, determining whether an individual is "Addict" or "Non-Addict." Ultimately, the final classifications are evaluated using a confusion matrix.

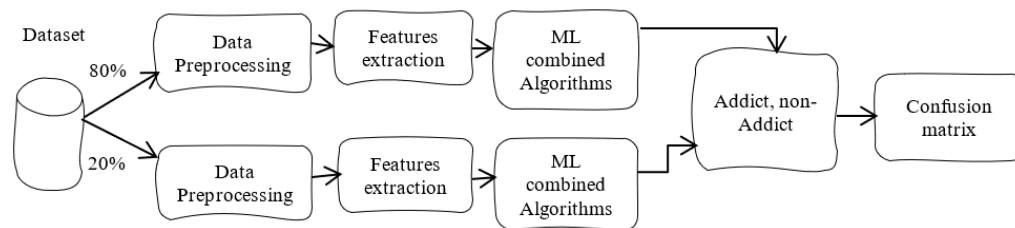


Figure 2. Data splitting and model evaluation

2.5.1. K-means

The K-Means algorithm is the most widely used technique for partitioning-based clustering. Patients were grouped into clusters based on proximity criteria using the K-means algorithm [29]. The algorithm is composed of the following steps: i) position K points within the space representing the patients to be clustered. These points act as the initial centroids of the groups, ii) assign each patient to the cluster with the nearest centroid, and iii) once all patients have been assigned to groups, recalculate the positions of the K centroids. Repeat steps 2 and 3 until the centroids remain stationary. This process divides the patients into homogenous groups while maximizing heterogeneity between the groups [30].

2.5.2. Linear regression

LR is utilized to estimate a linear hypothesis function between the output and input variables, serving as a regression or classification tool, and is expressed as [31]:

$$h_{\theta} = \theta_0 + \theta_1 \cdot X_1 + \dots + \theta_n \cdot X_n$$

Where h_{θ} is the hypothesis function, X_i represents input variables, and θ_i represents weights of the hypothesis function.

Weights represent the parameters of the hypothesis function. For estimating weight values, the first step is to calculate the error between the estimated result \hat{y} and the expected result (y) using the cost function. The mean squared error (MSE) is the most widely used cost function, and it is written as:

$$j = \frac{1}{N} \sum_{i=1}^N (\hat{y}(x^{(i)}) - y^i)^2$$

The second step involves using the gradient descent algorithm, which is the most crucial part of LR. This straightforward algorithm helps estimate the optimal weight values by minimizing the cost function. Gradient descent is an iterative process that updates the weights in each iteration to reduce the cost function, stopping when a threshold value is reached [32].

2.5.3. Evaluation criteria

To evaluate the effectiveness of our model, it is essential to use a range of performance metrics. Critical measures such as accuracy, sensitivity, specificity, and precision are vital for offering a detailed assessment of the model's performance. These metrics help identify the model's strengths and limitations in predicting outcomes, particularly in diverse scenarios or subgroups [33].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

3. RESULTS AND DISCUSSION

3.1. Descriptive analysis

3.1.1. Socio-demographic information

As shown in Table 1, cannabis addiction is much more prevalent among males in this dataset than females (98.6%). It is worth noting that the majority of the study participants are single (87%). Additionally, 60.2% of the addicts have an education level below high school. In contrast, the non-addicts group has an academic education level of 67.1% and a high school education level of 15.1%. Indeed, early cannabis use among adolescents is frequently linked to a decline in academic performance [7]. Also, among the addicts, 43.8% are unemployed and 56.2% are employed. While 87.7% of the non-addicts are unemployed. This difference is due to the fact that the non-addicted population is largely composed of students. Furthermore, the majority of participants come from an urban or suburban environment (97.2%), and most of them live in residential areas (76.7%). This sociodemographic result is consistent with a recent study conducted in Morocco exploring the relationship between cannabis use and worsening schizophrenia [34]. Additionally, the age distribution is similar between addicts and non-addicts, though addicts tend to be slightly older on average. The overall average age of all participants is 20.4 years, with a SD of 2.7 as shown in Table 2.

Table 1. Descriptive statistics for numerical data

| Feature | Addict | | | | Non-addict | | | | All | | | |
|---------|--------|------|-----|-----|------------|------|-----|-----|-------|------|-----|-----|
| | Means | SD | Min | Max | Means | SD | Min | Max | Means | SD | Min | Max |
| Age | 20.90 | 2.91 | 14 | 25 | 20 | 2.2 | 13 | 25 | 20.45 | 2.66 | 13 | 25 |
| AFCU | 15.47 | 2.89 | 2 | 22 | 3.85 | 6.60 | 0 | 19 | 9.66 | 7.73 | 0 | 22 |
| CCD | 3.95 | 2 | 1 | 10 | 0.0 | 0.00 | 0 | 0 | 1.97 | 2.45 | 0 | 10 |
| MoCA | 24.12 | 2.70 | 16 | 29 | 26 | 1.50 | 23 | 29 | 25.07 | 2.38 | 16 | 29 |
| PSQI | 9.60 | 3.43 | 2 | 19 | 6.53 | 3.1 | 0 | 18 | 8.06 | 3.59 | 0 | 19 |

3.1.2. Characteristic of cannabis addiction

Table 1 demonstrates that 28.8% of non-addict participants reported using cannabis in the past 12 months. Although this percentage is moderate, it is still notable for non-addicts, as occasional use can facilitate progression to cannabis dependence [35]. Furthermore, 46.6% of addicts reported that their parents used cannabis, while only 37% of non-addict users reported having a similar family background. This indicates a potential relationship between family cannabis use and an individual's dependence. This finding is supported by a study examining the impact of the family environment on adolescent cannabis use [36].

The CAST shows that 46.6% have a high risk of dependence, while only 2.7% have a low risk of dependence among non-addicts. In the same sense, the DSM-5 test reveals a clear disparity between addicts and non-addicts. The results indicate that 85% of addicts have moderate or mild dependence, while 15.1% have severe dependence. In contrast, only 2.7% of non-cannabis users have a mild dependence as presented in Table 2. It is important to highlight that both the CAST and the DSM-5 have good psychometric properties for cannabis screening and share many characteristics, but the CAST has not been validated against the DSM-5 criteria [37].

Concerning the age of first cannabis use (AFCU) and cannabis consumption duration (CCD) as shown in Table 2, cannabis addicts tend to have started using cannabis earlier than non-addicts, with a large difference in the AFCU. Indeed, the mean age of first use of cannabis among addicts is 15.47 years (SD=2.89), while the AFCU among non-addicts is 3.85 years (SD=6.60). Starting cannabis use at an early age can negatively impact an individual's mental health [38]. Also, CCD among addicts is 3.95 years (SD=2), while among non-addicts this information is not significant. It is essential to note that regular cannabis use is consistently linked to mild to moderate impairments in certain cognitive functions [24].

3.1.3. Cognitive status and sleep quality

Table 2 shows that the mean MoCA score of cannabis addicts is 24.12 (SD=2.70), while that of non-addicts is 25.10 (SD=1.50). This difference indicates a slight cognitive impairment in addicts, reflecting the effects of cannabis dependence on cognitive functioning, a widely recognized fact in the scientific literature [39]. Additionally, the mean PSQI score for addicts is 9.60 (SD=3.43), compared to 6.53 (SD=3.10) for non-addicts, indicating that addicts experience poorer sleep quality, as shown by their higher PSQI score, suggesting more significant sleep disturbances. Similar findings regarding cannabis use and poor sleep quality have been reported in studies involving adolescents [40].

3.1.4. Anxiety and depression levels and psychiatric profiles of patients and their families

Table 1 reveals that 41% of addicted participants had a psychiatric disorder, while only 1.4% of non-addicts reported similar conditions, highlighting a significantly higher prevalence of psychiatric disorders among the addicted group. These findings align with research on the relationship between cannabis use and

mental health, which consistently shows a high prevalence of cannabis use among individuals with mental illness [41]. Furthermore, 46.6% of addicts reported a family history of psychiatric disorders compared to only 13.7% of non-addicts, suggesting a possible genetic or environmental link between family psychiatric history and cannabis addiction [42].

Moreover, anxiety was reported by 49.3% of addicts, compared to 38.4% of non-addicts. Furthermore, 63% of addicts reported depression, compared to only 6.8% of non-addicts, demonstrating that depression is significantly more common among cannabis addicts. The HAD test shows that addicts have symptoms of depression and anxiety. In fact, 32.9% have definitive signs, and 46.6% have doubtful symptoms. In comparison with non-addicts, 69.9% have no symptoms of anxiety-depression, and 37% have doubtful symptoms. The HAD measurement reinforces the higher prevalence of mental health problems among cannabis addicts. Indeed, studies of early cannabis use in adolescents have demonstrated an increased risk of developing anxiety, depression, and psychosis [38].

Table 2. Descriptive statistics for categorical data

| Feature | Attribute description | N | % Addict | N | % Non-addict | N | % All |
|---|----------------------------|----|-------------|-----|-----------------|-----|----------|
| Sex | 0: Man | 72 | 98.6 | 37 | 50.7 | 109 | 74.7 |
| | 1: Women | 1 | 1.4 | 36 | 49.3 | 37 | 25.3 |
| Status | 0: single | 60 | 82.2 | 67 | 91.8 | 127 | 87 |
| | 1: married | 7 | 9.6 | 4 | 5.5 | 11 | 7.5 |
| | 2: divorced | 6 | 8.2 | 2 | 2.7 | 8 | 5.5 |
| Education | 0: Illiterate | 2 | 2.7 | 6 | 8.2 | 8 | 5.5 |
| | 1: Preschool | 2 | 2.7 | 1 | 1.4 | 3 | 2.1 |
| | 2: Primary | 13 | 17.8 | 1 | 1.4 | 14 | 9.6 |
| | 3: Secondary | 27 | 37.0 | 5 | 6.8 | 32 | 21.9 |
| | 4: High school | 15 | 20.5 | 11 | 15.1 | 26 | 17.8 |
| | 5: Academic | 14 | 19.2 | 49 | 67.1 | 63 | 43.2 |
| Profession | 0: no | 32 | 43.8 | 64 | 87.7 | 96 | 65.8 |
| | 1: yes | 41 | 56.2 | 9 | 12.3 | 50 | 34.2 |
| Settlement | 0: urban | 52 | 71.2 | 60 | 82.2 | 112 | 76.7 |
| | 1: suburban | 21 | 28.8 | 9 | 12.3 | 30 | 20.5 |
| | 2: rural | 00 | 00 | 4 | 5.5 | 4 | 2.7 |
| Neighborhood | 0: Popular | 14 | 19.2 | 20 | 27.4 | 34 | 23.3 |
| | 1: Residential | 59 | 80.8 | 53 | 72.6 | 112 | 76.7 |
| Use of cannabis in the last 12 months (CU12M) | 0: no | 73 | 100 | 52 | 71.2 | 52 | 35.6 |
| | 1: yes | 00 | 00 | 21 | 28.8 | 94 | 64.4 |
| Psychiatric disorder (PD) | 0: no | 43 | 58.9 | 72 | 98.6 | 115 | 78.8 |
| | 1: yes | 30 | 41.4 | 1 | 1.4 | 31 | 21.2 |
| Family psychiatric disorder attribute (FPDA) | 0: no | 28 | 53.3 | 63 | 86.3 | 91 | 62.3 |
| | 1: yes | 45 | 46.6 | 10 | 13.7 | 55 | 37.7 |
| Family cannabis user attribute (FCUA) | 0: no | 39 | 53.4 | 46 | 63 | 85 | 58.2 |
| | 1: yes | 24 | 46.6 | 27 | 37 | 61 | 41.8 |
| Psychometric properties of the cannabis abuse screening test (CAST) | 0: No risk of dependence | 1 | 1.4 | 71 | 97.3 | 72 | 49.3 |
| | 1: Low risk of dependence | 38 | 52.2 | 2 | 2.7 | 40 | 27.4 |
| | 2: High risk of dependence | 34 | 46.6 | 0 | 0 | 34 | 23.3 |
| Addictive profile in the last 12 months according to DSM-5 criteria (DSM-5) | 0: No addiction | 00 | 00 | 71 | 97.3 | 71 | 48.6 |
| | 1: Mild addiction | 18 | 24.760.3 | 2 | 2.7 | 20 | 13.7 |
| | 2: Moderate addiction | 44 | 15.1 | 0 | 0 | 44 | 30.1 |
| | 3: Severe addiction | 11 | | 0 | 0 | 11 | 7.5 |
| Anxiety | 0: no | 37 | 50.7 | 45 | 61.6 | 82 | 56.2 |
| | 1: yes | 36 | 49.3 | 28 | 38.4 | 64 | 43.8 |
| Depression | 0: no | 27 | 37 | 68 | 93.2 | 95 | 65.1 |
| | 1: yes | 46 | 63 | 5 | 6.8 | 51 | 34.9 |
| HAD | 0: absence of symptoms | 15 | 20.5 | 43 | 68.9 | 58 | 39.7 |
| | 1: doubtful symptoms | 34 | 46.6 | 273 | 37 | 61 | 41.8 |
| | 2: definite symptomatology | 24 | 32.9 | | 4.1 | 27 | 18.5 |

3.2. Correlational analysis

In order to enhance our comprehension of the dataset and the interrelationships among its attributes, we created a heatmap that illustrates the pairwise correlations between the 21 features that were employed in this study (refer to Figure 3). Based on the correlation value of each cell, the heatmap displays a color scheme that ranges from strongly positive (dark colors) to highly negative (light shades) associations. Strong positive relationships are shown by correlation values close to 1, and strong negative relationships are indicated by values close to -1. A correlation score close to 0, on the other hand, denotes little to no linear association between the corresponding features [43].

The heatmap shows a strong positive correlation between the 'Diagnostic' and 'CAST', and 'diagnostic' and 'DSM-5' features. This is an explainable fact since CAST and DSM5 are tests to assess the harmful use of cannabis or the severity of addiction. Other strong positive correlations are apparent in the heatmap, for example, the strong correlation between "Diagnosis" and "CU12MP", "Diagnosis" and "AFCU", and "Diagnosis" and "CCD". The heatmap helps evaluate potential multicollinearity among the features. Finding multicollinearity is important since it affects how well predictive models work and how comprehensible they are.

3.3. Predictive analysis

The elbow method is employed to determine the optimal number of clusters. This method evaluates the mean distance of observations to their respective centroids. As the number of clusters k increases, the intra-cluster variance decreases. A smaller intra-cluster distance is preferable, as it indicates more compact clusters. The elbow method identifies the value of k at which further increases do not significantly enhance the mean intra-cluster distance. Elbow method (Figure 4) illustrates optimal numbers of cluster size identified by using the Elbow method. The graph shows the possible optimal number of clusters $K=2$.

Figure 5 presents scatter plots with regression lines for K-means clusters, highlighting the relationship between DSM-5 scores and CAST scores. It consists of three subplots: Figure 5(a) exhibits the non-addict K-means cluster with a slope of 0.99 and an intercept of 0.0244. This plot shows a strong positive linear relationship, where DSM-5 and CAST scores are nearly proportional, meaning the CAST score increases almost equally with each unit increase in DSM-5. The intercept close to zero indicates that when DSM-5 is zero, the CAST score is nearly zero, which aligns well with the data. Figure 5(b) represents the addict K-means cluster with a slope of 0.70 and an intercept of 0.1168. This plot reveals a moderately positive linear relationship, where each unit increase in DSM-5 corresponds to a 0.7 unit increase in the CAST score. Although the relationship is positive, it is less steep compared to the non-addict group. Figure 5(c) combines both groups (addict and non-addict) and their regression lines. The distinct separation of the regression lines suggests that using the combined algorithm (K-means clustering followed by LR) effectively differentiates between addicts and non-addicts based on DSM-5 and CAST scores. It's important to highlight that the selection of CAST and DSM-5 features for LR is driven by the strong correlation shown in the heatmap in Figure 3.

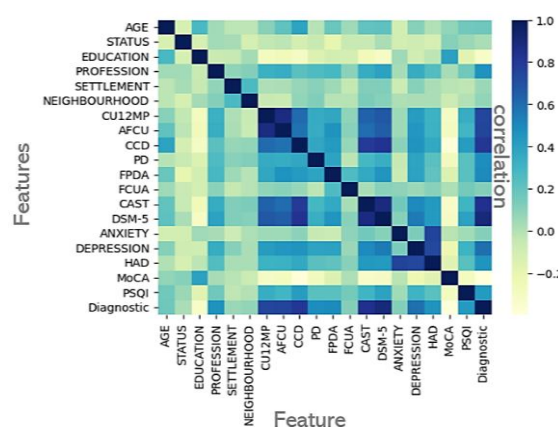


Figure 3. Data features correlation heatmap

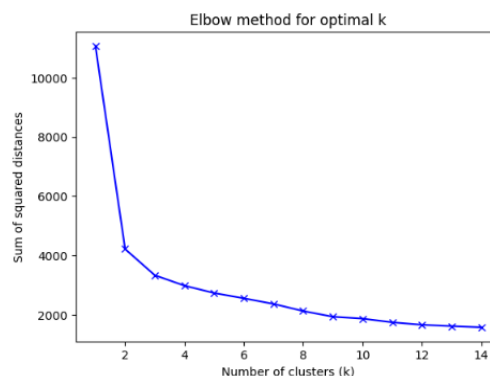


Figure 4. Elbow method: optimal number of clusters

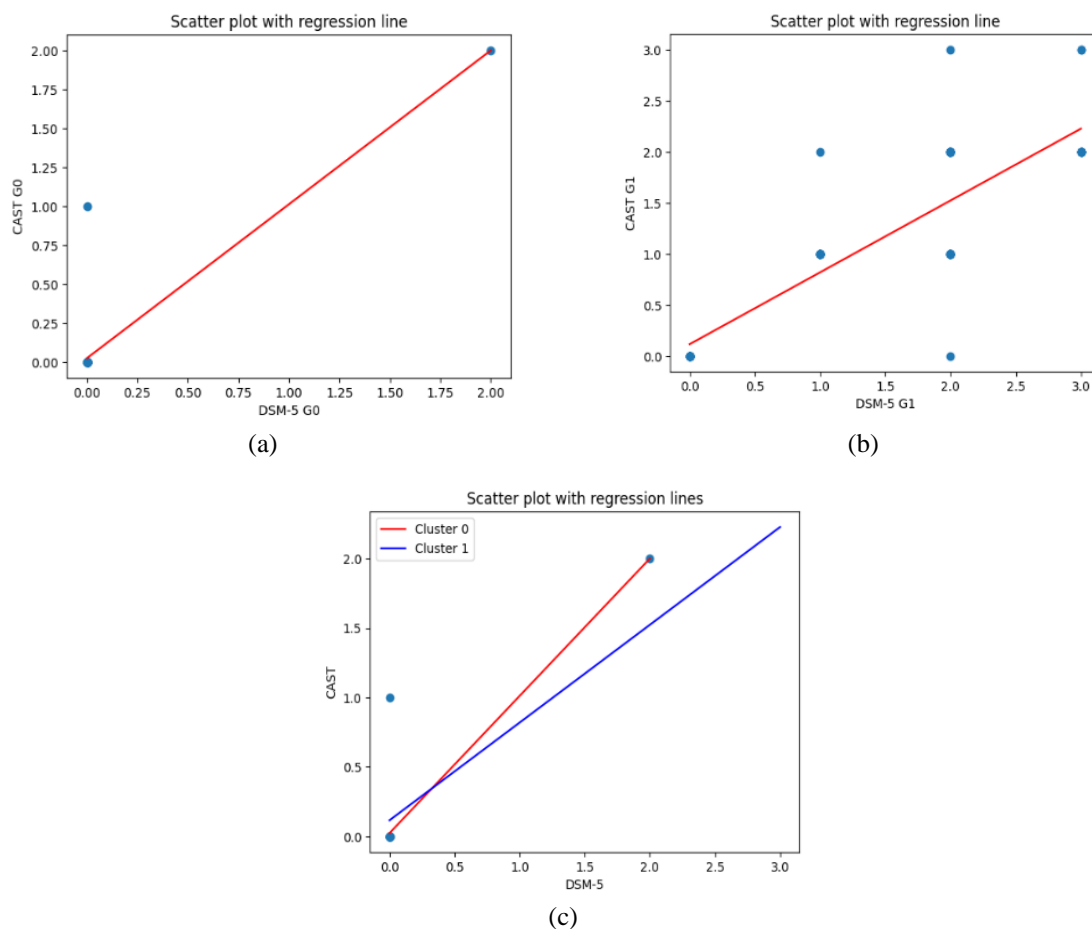


Figure 5. LR of K-means clusters of (a) LR for the non-addict clusters, (b) LR for the addict clusters, and (c) LR for the non-addict and addict clusters

The effectiveness of the proposed method was evaluated using sensitivity, specificity, and accuracy as performance metrics. The findings indicated that the combined algorithm reached a satisfactory overall prediction accuracy of 70%. It also demonstrated acceptable sensitivity at 65% and strong specificity at 86% when comparing these outcomes with empirical research using ML for prediction in addiction studies. In fact, the K-means algorithm demonstrated a sensitivity of 0.63 and an accuracy of 0.54 in a study using educational data mining to predict alcohol and drug dependence among students. These metrics are notably lower compared to the performance achieved by our combined model [44]. Besides, Rajapaksha *et al.* [45] underscored the importance of predicting the risk of CUD amid the growing legalization of cannabis. They examined data from 94 regular cannabis users using a least absolute shrinkage and selection operator (LASSO) logistic regression model, which pinpointed seven key risk factors: age, enjoyment of initial cigarette smoking, impulsive sensation-seeking scale score, cognitive instability, neuroticism personality trait score, openness personality trait score, and conscientiousness personality trait score. The model achieved an accuracy of 0.66 and an area under the ROC curve of 0.65. Furthermore, Choi *et al.* [46] focused on minimizing the negative health impacts of e-cigarette and hookah use among U.S. youth by creating cessation programs. This was done by identifying predictors of nicotine addiction and employing ML for predictive modeling. An analysis of data from 6,511 participants in the 2019 National Youth Tobacco Survey was conducted using random forest and LASSO methods. The resulting model, which incorporated 193 predictors, demonstrated notable accuracy, with LASSO scoring 0.6370 and random forest achieving 0.7342. Besides, another study investigates the traits of adult cannabis addicts (aged 20–49), with an emphasis on behavioral and social factors associated with depression and suicide risk. The prediction of depression risk uses baseline ML models, including logistic regression, random forest, and K-Nearest Neighbor. The performance of these models produced predictions similar to those found in our research [47]. It is evident that our combined algorithm's performance aligns with existing research, showing competitive accuracy and robust specificity. However, the lower sensitivity suggests a possible trade-off in correctly identifying all

positive cases, indicating areas that could benefit from improvement. Lastly, the combined algorithm used in this study for predicting addiction status yields promising outcomes, exhibiting competitive accuracy and high specificity. Ongoing efforts to enhance sensitivity and validate the algorithm in addiction-related applications are crucial for further advancing predictive modeling in this area.

3.4. Study limitations

The data collected in this study reflects the socio-demographic structure of the Marrakech region. It is observed that there are more male participants than female ones, and individuals from urban and suburban areas are more represented than those from rural regions. These imbalances create a certain bias in the data, making it difficult to generalize the results to other locations or populations. To address these limitations, it is recommended to apply a similar methodology to a larger dataset, with a greater inclusion of female and rural participants. Additionally, it would be beneficial to implement this approach on local data from other regions in Morocco or globally, adjusting the ML model accordingly.

4. CONCLUSION





This research aimed to develop an accurate ML-based model for predicting and classifying cannabis addiction by combining K-means clustering with LR to distinguish between addicted and non-addicted individuals. To enhance addiction diagnosis, the study integrated multiple validated, user-friendly clinical tests into a single consultation. The proposed method was tested using data from a primary care addiction center, and the results demonstrated that integrating K-means with LR is effective in predicting cannabis addiction. The model is designed to be accessible even to non-practitioners, addressing challenges related to understaffing and assisting in the implementation of preventive measures. The use of ML for addiction prediction holds promise across various sectors. It could be particularly beneficial in recruitment processes within industries like transportation and education, where cannabis dependence could have serious, harmful, or negative outcomes. Furthermore, this method can also be used to predict cannabis dependence when it is consumed in conjunction with other substances such as opioids and alcohol. In addition, the proposed method can be adapted to predict the severity of dependence. By meeting the demand for efficient and user-friendly diagnostic tools, this research supports ongoing initiatives in public health and addiction treatment. Future research could investigate applying this model to larger or more diverse populations and consider its integration into policies designed to mitigate the harmful effects of addiction.

REFERENCES





- [1] UNODC, *Global overview: drug demand drug supply*. Vienna, Austria: United Nations publication, 2021.
- [2] N. Gukasyan and E. C. Strain, "Relationship between cannabis use frequency and major depressive disorder in adolescents: findings from the national survey on drug use and health 2012–2017," *Drug and Alcohol Dependence*, vol. 208, no. September 2019, p. 107867, 2020, doi: 10.1016/j.drugalcdep.2020.107867.
- [3] W. Hall, "The costs and benefits of cannabis control policies," *Dialogues in Clinical Neuroscience*, vol. 22, no. 3, pp. 281–287, Sep. 2020, doi: 10.31887/DCNS.2020.22.3/whall.
- [4] UNODC, *Drug and age: drugs and associated issues among young and older people*, vol. 14, no. 3. Vienna, Austria: United Nations publication, 2018.
- [5] N. Castellanos-Ryan, J.-B. Pingault, S. Parent, F. Vitaro, R. E. Tremblay, and J. R. Séguin, "Adolescent cannabis use, change in neurocognitive function, and high-school graduation: a longitudinal study from early adolescence to young adulthood," *Development and Psychopathology*, vol. 29, no. 4, pp. 1253–1266, Oct. 2017, doi: 10.1017/S0954579416001280.
- [6] J. Wilson, T. P. Freeman, and C. J. Mackie, "Effects of increasing cannabis potency on adolescent health," *The Lancet Child & Adolescent Health*, vol. 3, no. 2, pp. 121–128, Feb. 2019, doi: 10.1016/S2352-4642(18)30342-0.
- [7] J. Leung, W. Hall, and L. Degenhardt, "Adolescent cannabis use disorders," in *Adolescent Addiction*, Second Edi., Elsevier, 2020, pp. 111–135. doi: 10.1016/B978-0-12-818626-8.00004-9.
- [8] E. L. Karschner, M. J. Swortwood-Gates, and M. A. Huestis, "Identifying and quantifying cannabinoids in biological matrices in the medical and legal cannabis era," *Clinical Chemistry*, vol. 66, no. 7, pp. 888–914, Jul. 2020, doi: 10.1093/clinchem/hvaa113.
- [9] E. Navarro-Tapia, J. Codina, V. J. Villanueva-Blasco, Ó. García-Algar, and V. Andreu-Fernández, "Detection of the synthetic cannabinoids AB-CHMINACA, ADB-CHMINACA, MDMB-CHMICA, and 5F-MDMB-PINACA in biological matrices: A systematic review," *Biology*, vol. 11, no. 796, pp. 1–14, 2022, doi: 10.3390/biology11050796.
- [10] H. E. Malki *et al.*, "Psychometric properties of the cannabis abuse screening test (CAST) in a sample of Moroccans with cannabis use," *Addiction Science & Clinical Practice*, vol. 19, no. 24, pp. 1–10, 2024, doi: 10.1186/s13722-024-00459-5.
- [11] N. Hemsing and L. Greaves, "Gender norms, roles and relations and cannabis-use patterns: a scoping review," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, p. 947, Feb. 2020, doi: 10.3390/ijerph17030947.
- [12] M. A. Schuckit, D. F. Clarke, T. L. Smith, and L. A. Mendoza, "Characteristics associated with denial of problem drinking among two generations of individuals with alcohol use disorders," *Drug and Alcohol Dependence*, vol. 217, no. August, p. 108274, Dec. 2020, doi: 10.1016/j.drugalcdep.2020.108274.
- [13] D. Wiljer and Z. Hakim, "Developing an artificial intelligence-enabled health care practice: rewiring health care professions for better care," *Journal of Medical Imaging and Radiation Sciences*, vol. 50, no. 4, pp. S8–S14, Dec. 2019, doi: 10.1016/j.jmir.2019.09.010.
- [14] K. K. Mak, K. Lee, and C. Park, "Applications of machine learning in addiction studies: a systematic review," *Psychiatry Research*, vol. 275, no. March, pp. 53–60, 2019, doi: 10.1016/j.psychres.2019.03.001.
- [15] J. L. Gowin, P. Manza, V. A. Ramchandani, and N. D. Volkow, "Neuropsychosocial markers of binge drinking in young adults," *Molecular Psychiatry*, vol. 26, no. 9, pp. 4931–4943, Sep. 2021, doi: 10.1038/s41380-020-0771-z.

- [16] X. Dong *et al.*, “Machine learning based opioid overdose prediction using electronic health records,” *AMIA Annual Symposium Proceedings Archive*, vol. 2019, pp. 389–398, 2019.
- [17] P. Cresta Morgado, M. Carusso, L. Alonso Alemany, and L. Acion, “Practical foundations of machine learning for addiction research. Part I. methods and techniques,” *American Journal of Drug and Alcohol Abuse*, vol. 48, no. 3, pp. 260–271, 2022, doi: 10.1080/00952990.2021.1995739.
- [18] Y. Jing *et al.*, “Analysis of substance use and its outcomes by machine learning I. childhood evaluation of liability to substance use disorder,” *Drug and Alcohol Dependence*, vol. 206, no. February 2019, p. 107605, Jan. 2020, doi: 10.1016/j.drugalcdep.2019.107605.
- [19] Z. Che, J. St Sauver, H. Liu, and Y. Liu, “Deep learning solutions for classifying patients on opioid use,” *Annual Symposium proceedings (AMIA). AMIA Symposium*, vol. 2017, pp. 525–534, 2017.
- [20] U. Sakoglu, M. Mete, J. Esquivel, K. Rubia, R. Briggs, and B. Adinoff, “Classification of cocaine-dependent participants with dynamic functional connectivity from functional magnetic resonance imaging data,” *Journal of Neuroscience Research*, vol. 97, no. 7, pp. 790–803, Jul. 2019, doi: 10.1002/jnr.24421.
- [21] M. Liu *et al.*, “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use,” *Nature Genetics*, vol. 51, no. 2, pp. 237–244, Feb. 2019, doi: 10.1038/s41588-018-0307-5.
- [22] A. Elhachimi, A. Benksim, H. Ibanni, and M. Cherkaoui, “The screening of cannabis addiction using machine learning, moca, and anxiety/depression tests,” *Scientific African*, vol. 24, no. March, p. e02225, 2024, doi: 10.1016/j.sciaf.2024.e02225.
- [23] J. P. Connor, D. Stjepanović, B. Le Foll, E. Hoch, A. J. Budney, and W. D. Hall, “Cannabis use and cannabis use disorder,” *Nature Reviews Disease Primers*, vol. 7, no. 1, pp. 1–24, 2021, doi: 10.1038/s41572-021-00247-4.
- [24] M. E. Lovell, J. Akhurst, C. Padgett, M. I. Garry, and A. Matthews, “Cognitive outcomes associated with long-term, regular, recreational cannabis use in adults: a meta-analysis,” *Experimental and Clinical Psychopharmacology*, vol. 28, no. 4, pp. 471–494, Aug. 2020, doi: 10.1037/pha0000326.
- [25] H. El Malki *et al.*, “Psychometric properties of the cannabis abuse screening test (CAST) in a sample of moroccans with cannabis use,” *Addiction Science & Clinical Practice*, vol. 19, no. 1, p. 24, Apr. 2024, doi: 10.1186/s13722-024-00459-5.
- [26] M. S. Hamrah *et al.*, “Anxiety and depression among hyperlipidemic outpatients in afghanistan: a cross-sectional study in Andkhoy City,” *International Journal of Hypertension*, vol. 2018, pp. 1–8, Aug. 2018, doi: 10.1155/2018/8560835.
- [27] E. A. Winiger, L. N. Hitchcock, A. D. Bryan, and L. C. Bidwell, “Cannabis use and sleep: expectations, outcomes, and the role of age,” *Addictive Behaviors*, vol. 112, no. September 2020, p. 106642, Jan. 2021, doi: 10.1016/j.addbeh.2020.106642.
- [28] D. A. Anggoro and D. Permatasari, “Performance comparison of the kernels of support vector machine algorithm for diabetes mellitus classification,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, pp. 580–585, 2023, doi: 10.14569/IJACSA.2023.0140163.
- [29] M. Khalid, N. Pal, and K. Arora, “Clustering of image data using K-means and fuzzy K-means,” *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 7, pp. 160–163, 2014, doi: 10.14569/IJACSA.2014.050724.
- [30] C. Violán *et al.*, “Multimorbidity patterns with k-means nonhierarchical cluster analysis,” *BMC Family Practice*, vol. 19, no. 1, pp. 1–11, 2018, doi: 10.1186/s12875-018-0790-x.
- [31] K. Bazdaric, D. Sverko, I. Salaric, A. Martinović, and M. Lucijanic, “The ABC of linear regression analysis: what every author and editor should know,” *European Science Editing*, vol. 47, pp. 0–9, Sep. 2021, doi: 10.3897/ese.2021.e63780.
- [32] T. M. H. Hope, “Linear regression,” in *Machine Learning: Methods and Applications to Brain Disorders*, Elsevier, 2020, pp. 67–81, doi: 10.1016/B978-0-12-815739-8.00004-3.
- [33] M. E. A. Bourkha, A. Hatim, D. Nasir, S. El Beid, and A. S. Tahiri, “A novel inter patient ecg arrhythmia classification approach with deep feature extraction and 1D convolutional neural network,” *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, p. 5, 2024, doi: 10.14569/IJACSA.2024.0150265.
- [34] E. H. Ouanouche *et al.*, “Cannabis and schizophrenia: characterisation of a risk factor in a sample of moroccan patients hospitalised for psychosis,” *Middle East Current Psychiatry*, vol. 29, no. 1, p. 12, Dec. 2022, doi: 10.1186/s43045-022-00173-5.
- [35] P. George and M. Wahl, “Cannabis toxicity in children and adolescents,” *Pediatric Annals*, vol. 52, no. 5, pp. e181–e186, 2023, doi: 10.3928/19382359-20230307-04.
- [36] M. S. Schuler, J. S. Tucker, E. R. Pedersen, and E. J. D’Amico, “Relative influence of perceived peer and family substance use on adolescent alcohol, cigarette, and marijuana use across middle and high school,” *Addictive Behaviors*, vol. 88, no. 1, pp. 99–105, Jan. 2019, doi: 10.1016/j.addbeh.2018.08.025.
- [37] S. Legleye, “The cannabis abuse screening test and the DSM-5 in the general population: optimal thresholds and underlying common structure using multiple factor analysis,” *International Journal of Methods in Psychiatric Research*, vol. 27, no. 2, pp. 1–10, Jun. 2018, doi: 10.1002/mpr.1597.
- [38] S. Hosseini and M. Oremus, “The effect of age of initiation of cannabis use on psychosis, depression, and anxiety among youth under 25 years,” *The Canadian Journal of Psychiatry*, vol. 64, no. 5, pp. 304–312, May 2019, doi: 10.1177/0706743718809339.
- [39] E. Cyrus *et al.*, “A review investigating the relationship between cannabis use and adolescent cognitive functioning,” *Current Opinion in Psychology*, vol. 38, no. July 2020, pp. 38–48, Apr. 2021, doi: 10.1016/j.copsyc.2020.07.006.
- [40] R. P. Ogeil *et al.*, “Early adolescent drinking and cannabis use predicts later sleep-quality problems,” *Psychology of Addictive Behaviors*, vol. 33, no. 3, pp. 266–273, 2019, doi: 10.1037/adb0000453.
- [41] S. A. Al Azizi, A. A. Omer, and A. A. Mufaddel, “Cannabis use among people with mental illness: clinical and socio-demographic characteristics,” *Open Journal of Psychiatry*, vol. 08, no. 03, pp. 244–252, 2018, doi: 10.4236/ojpsych.2018.83021.
- [42] C. R. Quick, K. P. Conway, J. Swendsen, E. K. Stapp, L. Cui, K. R. Merikangas, “Comorbidity and coaggregation of major depressive disorder and bipolar disorder and cannabis use disorder in a controlled family study,” *JAMA Psychiatry*, 79, no. 7, pp. 727–735, 2022, doi: 10.1001/jamapsychiatry.2022.1338.
- [43] K. Omari, “Comparative study of machine learning algorithms for phishing website detection,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, pp. 417–425, 2023, doi: 10.14569/IJACSA.2023.0140945.
- [44] Tanvi Trivedi, “Exploring prediction modeling of students alcohol and drug addiction affecting performance using data mining approach,” *International Journal of Engineering Research and*, vol. V8, no. 12, Dec. 2019, doi: 10.17577/IJERTV8IS120105.
- [45] R. M. D. S. Rajapaksha, R. Hammonds, F. Filbey, P. K. Choudhary, and S. Biswas, “A preliminary risk prediction model for cannabis use disorder,” *Preventive Medicine Reports*, vol. 20, p. 101228, Dec. 2020, doi: 10.1016/j.pmedr.2020.101228.
- [46] J. Choi, H. T. Jung, A. Ferrell, S. Woo, and L. Haddad, “Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (hookah) users,” *Journal of Clinical Medicine*, vol. 10, no. 5, pp. 1–13, Mar. 2021, doi: 10.3390/jcm10050972.
- [47] J. Choi, J. Chung, and J. Choi, “Exploring impact of marijuana (cannabis) abuse on adults using machine learning,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 19, p. 10357, Oct. 2021, doi: 10.3390/ijerph181910357.





BIOGRAPHIES OF AUTHORS

Abdelilah Elhachimi     is a permanent trainer at the Institute of Nursing Professions and Health Techniques of Marrakech (ISPITS), with basic training as a nurse anesthetist. In addition, he holds a master's degree in nursing education and another in neuroscience and biotechnology. Currently, he is pursuing doctoral research focused on the application of machine learning in the field of drug addiction studies. He can be contacted at email: abdelilah.elhachimi@ced.uca.ma.







Mohamed Eddabbah     born in 1986 in Marrakech City, Morocco, has a mixed academic and professional profile with a comprehensive background in Information and Communication Technology (ICT). Prior to his doctoral studies, he was a Telecommunications Engineer with a Ph.D. in Informatics, specializing in networks and telecommunications. Currently, he holds the position of Full-time Professor in Computer Science at the Higher School of Technology – Essaouira. Cadi Ayyad University, Marrakesh, Morocco. He significantly contributes to various aspects of computer science, including programming, software quality, software project management, computer security, IoT, digital transformation, and machine learning. He can be contacted at email: eddabbah@gmail.com or m.eddabbah@uca.ac.ma.







Abdelhafid Benksim     is a Full Professor affiliated with the Laboratory of Human Ecology in the Department of Biology at the School of Sciences Semlalia, Cadi Ayyad University in Marrakech, Morocco, and also with the Laboratory of Biology at the High Institute of Nursing and Health Techniques in Marrakesh, Morocco. He is currently the coordinator of a master's degree in emergency medicine in ISPITS. He can be contacted at email: benksima@gmail.com.



Hamid Ibannid     is a psychiatrist, child psychiatrist and addictologist, professor at the ISPITS Workers, regional manager of the Addictology Centers of Marrakech. Member of the board of the National Association of Drug-Risk Reduction DRR. in charge of prevention, support for recovery and reintegration of drug addicts. author of several works on addictology and psychiatry. He can be contacted at email: ibanni1968@gmail.com.



Mohamed Cherkaoui     is a professor and researcher at Cadi Ayyad University, associated with the Department of Biology. He focuses his research in the Neurosciences, Pharmacology, Anthropobiology, and Environment Laboratory. With extensive work in human ecology and biology, he is also an experienced statistician. He can be contacted at email: cherkaoui@uca.ac.ma.